IDIAP RESEARCH REPORT

IDIAP
Martigny - Valais - Suisse

A Survey of Text Detection and Recognition in Images and Videos

Datong Chen, Juergen Luettin, Kim Shearer

IDIAP–RR 00-38

August 2000

Institut Dalle Molle d'Intelligence Artificielle Perceptive • CP 592 • Martigny • Valais • Suisse

téléphone +41−27−721 77 11
télécopieur +41−27−721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

Rapport de recherche de l'IDIAP 00-38

A Survey of Text Detection and Recognition in Images and Videos

Datong Chen, Juergen Luettin, Kim Shearer

August 2000

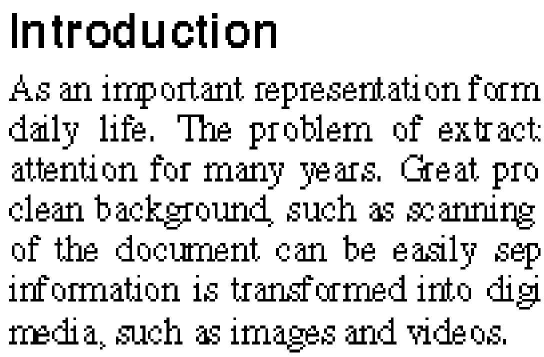

Introduction
As an important representation form
daily life. The problem of extract
attention for many years. Great pro
clean background, such as scanning
of the document can be easily sep
information is transformed into digi
media, such as images and videos.

FIG. 1 – *Text in document*

1 Introduction

As an important representation form of human beings language, visual texts are widely used in our daily life. The problem of extracting text information from visual clues have attracted wide attention for many years. Great progress has been made in processing printed characters against clean background, such as scanning document pages. The pixels of the text in the scanning result of the document can be easily separated from the background. While, today, more and more information is transformed into digital form, visual texts are embedded in many forms of digital media, such as images and videos.

Comparing with the texts in documents, the text in media is in small quantity, but often carries crucial information of the media contents. They usually present important names, locations, brands of the products, scores of the match, date and time, which are helpful information to understand and index these images and videos. However, the text in the images and videos can be superimposed on the arbitrary backgrounds or embedded on the surfaces of the objects in the scene with vary font, size, color, alignment, movement and lighting condition, which makes the text extraction extremely difficult. The aim of research on text detection and recognition in images and videos focuses on finding the proper way to extract different types of text from arbitrary complex images or videos. It will not only extend the application of OCR system into wider multimedia areas but also help people further understand the mechanism of the visual text detection and recognition.

1.1 Definitions of the terms of text retrieval in images and videos

To efficiently introduce the research works on text detection and recognition in images and videos, it is necessary to clearly define some important terms we will use in this paper.

Document text: this term indicates the the text characters against clean background such as some scanning document pages. See

Media text: we use this term to describe the text in arbitrary images and videos. Comparing with the document text, the contents of the images and videos can be outdoor or indoor natural scenes, artificial graphs and pictures, even hybrid visual media.

Media Text Detection: to answer whether there are texts in the image or specified period of video stream.

Media Text Location: simply, it should output the exact position of the regions of the text. If the text region is rectangle, its position can be easily represented. Since it is not easy to define the position of the arbitrary-shape region, in this paper, the term text location only indicates to locate the text region in a rectangle box.

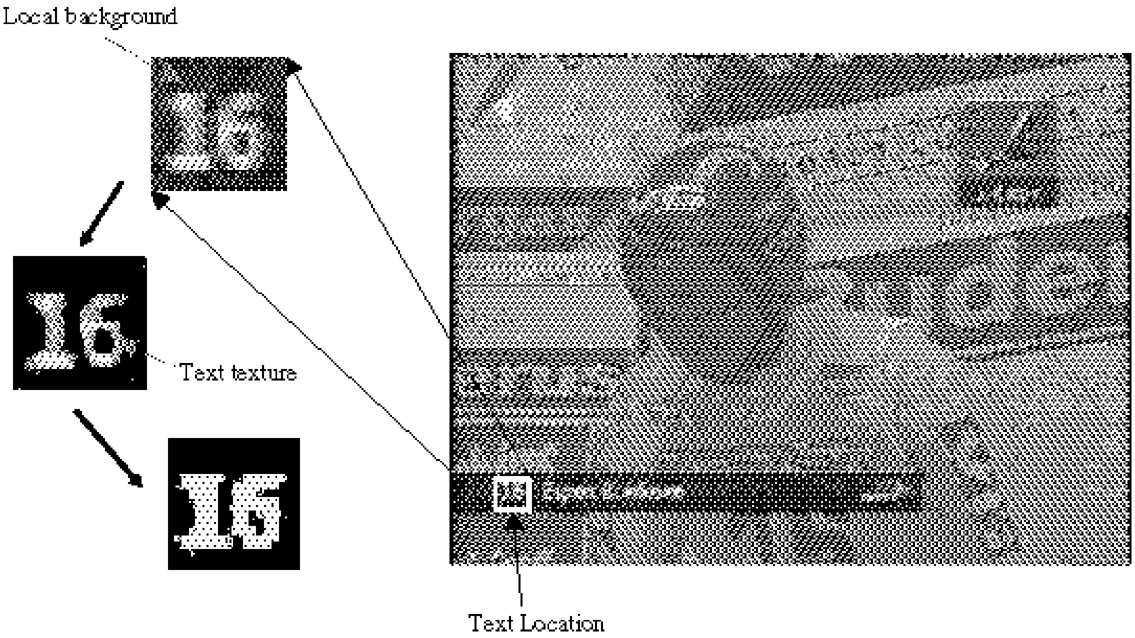


FIG. 2 – *Text in the image/video*

Text texture: the texture consists of the pixels belong ot the media text.

Local background: indicates certain neighborhood area of the text texture. We introduce this term because there are a lot text embedded in the images or videos with surrounded artificial background.

Media Text segmentation: to separate the text texture from other pixels in the image or video.

Media Text recognition: convert the distinguishable text texture in different background into uniform coded text.

These definitions are only used to describe the related research work in this paper. We hope that they can be elaborated to introduce a common terminology and a more clear definition of the addressed problems. This will facilitate the exchange and comparison of results with other research groups.

1.2 Application of the Media Text Recognition Research

Today, more and more information is stored in vary kinds of digital forms including the images and videos. Through the broadcasting, laser storage material and internet, digital media leads our daily life into a new age while brings us some new challenges. One general problem is the management of such vast information sources. The problems about how to organize, store, search, and use the multimedia activate the content-based image and video/audio analysis research in recent years, which offers the media text recognition a wide application area.

1.2.1 Text-based images and video indexing

Efficient descriptions are needed for indexing and retrieving the images and videos. The low level feature based content representation, such as color and texture, is difficult for user to input as the keywords for indexing and retrieving while the index and retrieve methods of these low level features are difficult to perform as efficiently as the text-based search engines. Most of high level contents representations, such as human face and body, physic objects, activities, in images and videos are not only difficult to be retrieved automatically, but also difficult to be described and matched. The image/video-based indexing and retrieval systems need the efficient descriptors to enjoy their own advances. As a form of high level visual content text provides more precise and explicit meaning and easier description than low-level and other high level visual features. There are two facts to show the comfortable for indexing and retrieving by using text-based search engine and OCR system.

1. A powerful source of information is the text contained in images and videos. Superimposed captions in news videos usually provide information about the name of related people, location, subject, date and time. Captions also often provide an abstract or summary of the program. This information is often not available in other media like audio "track". Titles and credits displayed at the beginning or end of movies provide information like names of actors, producers, contributors, etc. Captions in sport programs often contain the names of teams, players, scores, etc. Text information can also be found in the scene, e.g. the players numbers and name, the name of the team, brand names, location, and commercials. Displayed maps, figures and tables in videos contain much text about locations, temperatures, certificate items. Titles, logos, and names of programs displayed in video are important for the annotation with respect to program types and names. Even the licenses and brand names of the vehicles, conversion of the paper documents, like covers of CDs, books, and journals, in a video stream or image contain valuable information. More and more web sites choose pictures to improve the visual effect of their titles. These media texts provide therefore an important source for indexing, annotation and other content oriented processing.

2. Both the text-based searching and matching technologies are well developed. Text-based search engines have been well developed to address the problem of automatically and efficiently finding documents. Many efficient search engines are employed in peer visiting web sites, e. g. Yahoo and Alta Vista, libraries, and a lot of business areas. Document OCR systems for machine printed text on clean papers yield high recognition rates and are now readily available for personal computers [3][4]; see [5] for an overview of methods and current research issues.

Combining the text-based search engine and the OCR technologies seems a good solution for the images and videos indexing and retrieving. However, the document OCR systems usually require high resolution and contrast, and uniform background with simple gray value distribute text as input. Current optical character recognition technologies are not able to recognize text that is not printed against a clean background. Similarly, document analysis methods that try to segment whole pages into individual segments usually require a binary input and assume a specific document layout, e.g. newspaper or technical article. In images and videos, the background can be any kinds of indoor or outdoor scenes while the text can vary in font, size, color, texture, alignment, 3D position and movement, lighting condition, and shading. Therefore, the current OCR can not directly be applied for the recognition of text in images or videos. There is thus a high demand for systems that find, extract, and recognize unconstrained text from any background.

1.2.2 Content-oriented video coding

Beside indexing and retrieval, another application of media text detection and recognition can be found in content oriented multimedia coding. Since the text carries clear information to the observers, text texture should be carefully processed when the image and video are compressed. The text texture can be extracted from the images and video as a kind of visual object and encoded with special algorithms to achieve higher compression rate or image quality. Furthermore, the text object can be recognized into text and applied in MPEG-4 SNHC (synthetic and natural hybrid coding) with proper synthesis algorithms. Media text detection and recognition offers a opportunity to extract the text video object.

1.2.3 Application on internet and digital library

Moreover, the applications of the media OCR can be found in the industry of the internet and the digital libraries. The typical applications in these areas include:

– transcription and acoustic output of text for blind WWW users.

- conversion of documents into electronic versions.
- conversion of paper-based advertisement into electronic form to automatically generate its electronic equivalent for WWW publication.

1.3 Scientific validation

Different with the document text processing, the research on media text detection addresses the problem about how to extract the text patterns from the arbitrary background. Evidence shows that the human vision system is so sensitive to the texture produced by visual text that text can be extracted accurately even when the texts are in different size, font, gray level and color, opaque, partially occluded. Do we know how to build a system to perform the similar thing?

Since the text texture has no special gray level or color distribution as the other objects do, for example the human face. Some opaque texts are represented by only adding certain gray or color value on the background directly, which partially enjoy the similar gray or color distribution with its background. The research on media text detection will try to feature the text patterns without clear gray or color distribution knowledge to distinguish the text from other objects.

The text in images and video may have varing sizes and fonts. Different sizes and fonts text attracts different attentions of the observer, while offering a huge searching space for text detection algorithm. Whether can we simply magnify the images or video frames to match the size of the standard patterns just as what happens in document text process? How to measure the efficiency of the searching? How to search vary sizes and fonts of text efficiently?

Is it possible to segment the texts without recognize them? In other words, whether there is a kind of common features of the text and non-text, which can distinguish the text texture from non-text texture and can not recognize each single character of the text? Proper investigation needs to be done to answer this question.

Usually, the images and videos consists of wider range of scenes than the document pages. The research on media text detection and recognition will add our knowledge about not only the character recognition, but also the special pattern, like text pattern, universal visual detection methods.

2 Architecture of Media Text OCR System

2.1 Classification of Media Text

2.1.1 According to the source of the text

The main types of text in images can be roughly classified into scene text and superimposed text according to the source of the text. Scene text, is text that is part of the scene whereas superimposed text is laid over the scene, e.g. by a video title machine.

Scene text can originate from a recorded scene or photograph. Examples of scene texts are name on T-shirt, commercials in sport stadiums, traffic signs. This kind of text, can either bear important information for indexing (city sign, name of player) or be rather incidental (commercials) and unsuitable for indexing. This kind of text is more difficult to detect, extract and recognize due to the nature of the scene, e.g. movement, lighting, affine transformation, and occlusion.

Superimposed text on the other hand generally contains relevant information. (Some researchers like to use the term "graphic text" to describe the same concept [1]). In the case of news, for example, it is usually generated to provide the viewer with key information about the current content of the program. This kind of text is therefore important for indexing. It is also often easier to extract since it appears under certain general constraints.

Scene text can be described by the following characteristics:

- Variability of size, font, color, orientation, style, alignment, even within words.
- Part of the text might be occluded.

- With complex movement in the scene.
- with variability of lighting conditions.
- with variability of transformations.
- deformation if on flexible surface (played shint). (Kim)

Superimposed text has usually the following characteristics:

- Text is always in the foreground and never foreground occluded.
- usually with stable lighting condition, which is scene-independent.
- The values of the text pixels are distributed according to limited rules.
- Size, font, spacing, and orientation are constant within a text region.
- Text is normally aligned horizontally.

The following additional characteristics are usually observed for superimposed text in video images:

- Text is either stationary or linearly moving in horizontal or vertical direction. (not exactly)
- The background is uniform for moving text. (not exactly)
- The same text appears in several consecutive frames.
- Low text resolution.

2.1.2 According to the properties in images and videos

In order to further illustrate the problems in dirty text recognition, a detail classification of the dirty text is made according to the properties of the text and the images and videos where the text exits. The dirty text should have at least one of the properties in each following column.

Since all the former research works in the related area based on some assumptions outlined above, this property-based classification is also helpful for us to understand them.

2.2 Architectures of former research

Since the clean text processing systems have lead rather good results, researchers hope to design the dirty text processing architecture basing on the existing technologies. There are many different ways to apply the exist OCR technologies on the dirty text recognition.

The simplest way is to clean the input images or videos with vary kinds of segmentation technologies so that they can be used as the input of the clean OCR system directly because it does not have to change any thing in the traditional OCR system. Unfortunately, most of the research works suggests that it is not possible to segment the pixels of the text without knowing where and what the characters are [35][27].

A tradeoff is to segment text-like texture instead of doing real text segmentation in the first step. And then the clean OCR technologies are employed to recognize the text from these text-like texture regions or reject them. The corresponding architecture of the OCR system is showed in figure 1. What is a text-like texture? Some researchers give out their own definitions. For example, [32] defined the text-like texture as a horizontal rectangular structure of clustered sharp edges. Other researchers prefer to implicitly define this concept in their assumptions or algorithms.

There are some differences between text segmentation in images and videos. One obvious difference is that the videos offer motion clues of the objects. We will discuss the typical architectures for both still images and videos.

TAB. 1 – *Media text classifications*

Category	Detail Properties	Source of the Text
Alignment of the Text	horizontal, strait line	Superimposed Text
	strait line in every directions	Superimposed & Scene Text
	curves	Superimposed & Scene Text
Local background of the Text	static artificial grapic	Superimposed Text
	static natural image	
	move artificial graphic	Superimposed Text
	move natural scene	
Movement of the Text	static	Superimposed Text
	liner movement	Superimposed Text
	2D rigid constrained movement	Superimposed & Scene Text
	3D rigid constrained movement	Superimposed & Scene Text
	free movement (with transformation)	Superimposed & Scene Text
Occlusion	text can not occluded	Superimposed Text
	text can be partly occluded	Superimposed & Scene Text
Transform of the text	no transformation	Superimposed Text
	3D rigid movement caused transform	Superimposed & Scene Text
	free transform	Superimposed & Scene Text
Size, font, color of the text	uniform within one word	Superimposed & Scene Text
	vary within one word	Superimposed & Scene Text
Resolution of the text	up than 40×40 pixels	Superimposed & Scene Text
	from 12×12 to 40×40 pixels	Superimposed & Scene Text
	less than 12×12 pixels	Superimposed Text
Contrast of the text and the local background	contrast is high at any parts of the text and at any time	Superimposed & Scene Text
	contrast is high at any parts of the text through a period of continues time	Superimposed & Scene Text
	partly low contrast (even in the color space)	Superimposed Text
	partly can not distinguish	Superimposed Text
Shading & other effects	different kinds of shading methods	Superimposed Text
	different kinds of Telop (Television Opaque Projection)	Superimposed Text

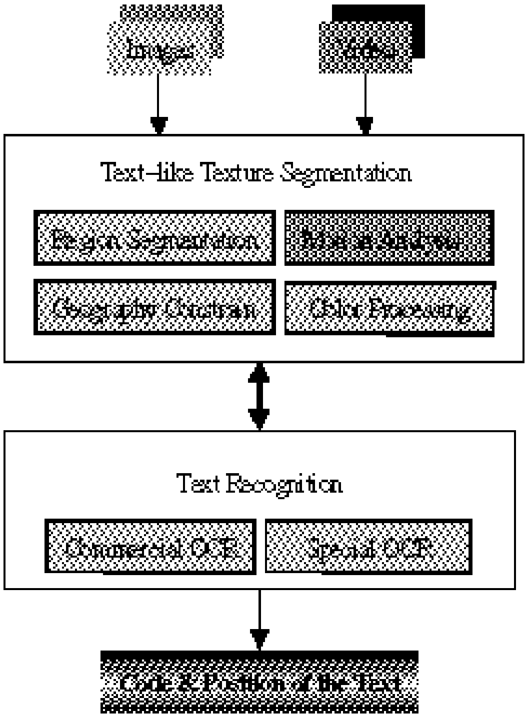


FIG. 3 – *Brief Architecture of Media OCR*

2.3 Text-like texture segmentation in images

The key point in designing text-like texture segmentation algorithm is to find a way to measure the difference between the text pixels and the "background" pixels. Due to the different definitions of the text-like texture, most of the available research work prefers designing measurement for each specific application rather than seeking the solution systematically. By examining the architectures of their algorithms, we hope to depict a clear picture about the ideas of these researches.

One of the early algorithm was presented by Ohya, which aims at extracting and recognizing the scene texts from images [12]. The algorithm consists of three steps. At the first step, the input images are roughly segmented into regions by using an image segmentation method based on adaptive thresholding. Second, the character candidate regions are selected through checking the features under the following assumptions:

- the gray-level of the text character region is high contrast to the background;
- the width of text character segmentation is uniform;
- gray-level of text character segmentation is uniform;
- spatial frequency of text character segmentation is uniform.

At the last step, the algorithm applies a recognition process to cluster the separate parts of one text character together and extract the character pattern candidates.

The applicable text characters of the text in this algorithm can be distorted in 3-D space under the uncontrolled illuminating conditions. Furthermore, the text characters can have varing sizes, pitches, positions, fonts, formats, and gray-levels.

From this three-step processes algorithm, we can clearly find three types of features are employed to measure the text-like texture in this algorithm. They are:

1. The statistic & distribution properties of the pixel values within the character and its local background, for example, the uniform gray-level of text character segmentation and high contrast to the background.

2. The 2D spatial distribution properties of the characters in a word or a sentence, such as the uniform spatial frequency of the text characters, text alignment.

3. The shape features of the characters, e. g. uniform width of the text character.

In Ohya's algorithm, the three kinds of features are used one by one to ensure the algorithm run fast. However, some tight constrains are used in the algorithm, such as the uniformity of gray-level, width and spatial frequency of the character segmentation, which limit the applicability of the algorithm.

Another typical algorithm of text segmentation in images is proposed by [36], which consisted of four phases. First, by assuming the text-like texture has certain statistic properties throughout the Gaussian scale space, a texture segmentation scheme is used to segment the images roughly into text regions and background. Second, the text regions are refined under the constraits coming from the assuming heuristics on text strings, such as height similarity (characters are the same size in a word or sentence), spacing and alignment (linear alignment with fix space between text lines). Third, the text region candidates are binarized according to the distribution properties of the pixel values. Finally, the text string candidates are refined by applying the same kind of constrains as used in the second phases with tighter standards. Different with Ohya's method, this algorithm focuses on locating the text string (words or sentence) but not the characters at the first step, which is also often used in most of other media text OCR systems.

Both of these two algorithms try to roughly locate the text region candidates at first without regarding the shape properties of the characters and then employ tighter constrains to identify and segment these candidates. There are some other algorithms for locating the text region in gray and color images by using texture analysis and connected component. [40]

The former works as we discussed above focus on detecting, locating or segmenting the text from the images with the complex background. Most of them assume that the text:

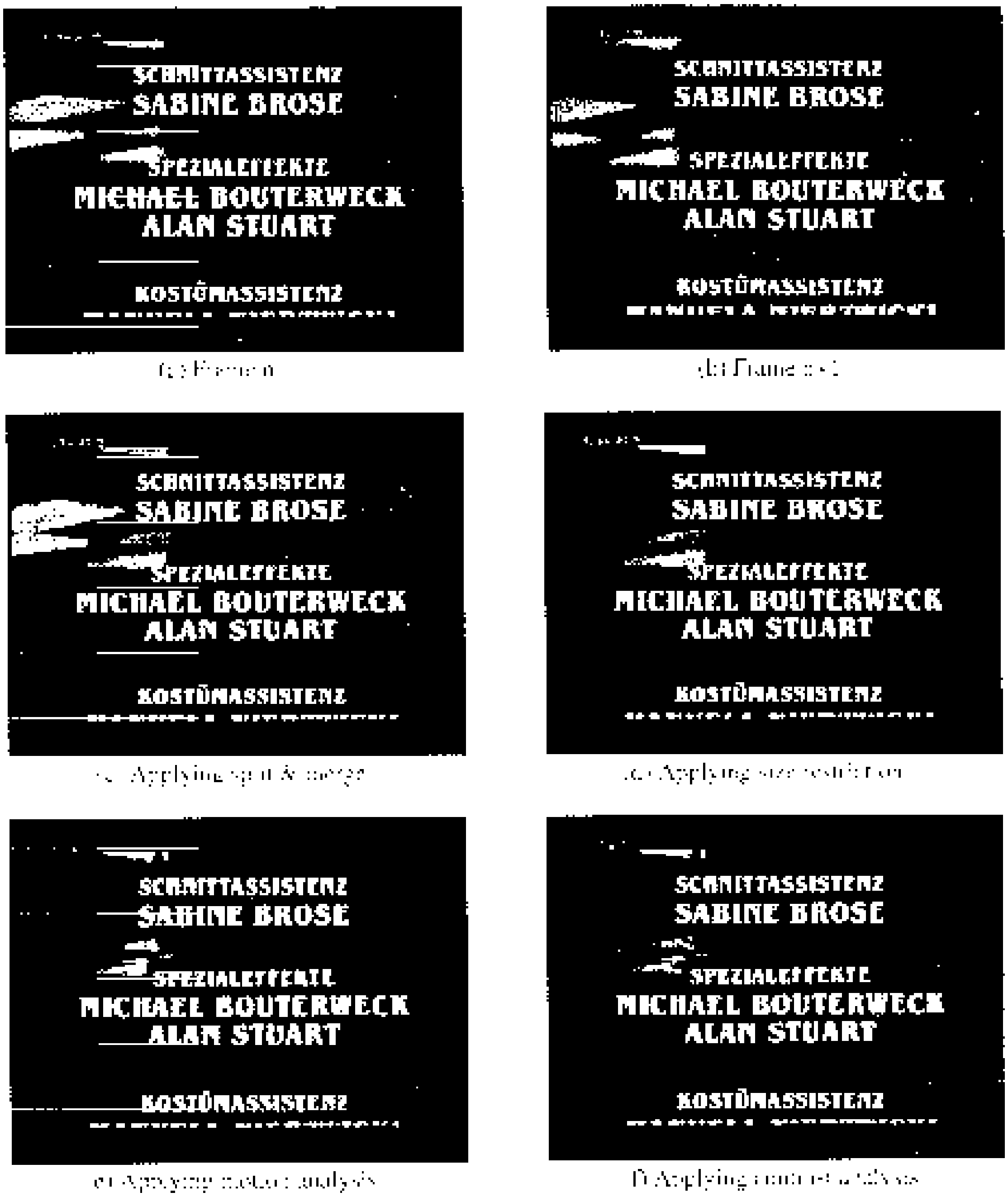
- may have vary size, but should keep the same font size with in the one word;
- should aligned in horizontal line. Some algorithms, for example the spatial variance method, can be extended to find the text in any linear alignment;
- should have high contrast with the local background in gray value or color value;
- can not be occluded.

Ohyas algorithm allow the text can have certain transformation. The OCR system build by themselves may tolerant this kind of transformation. There is no clear discussion about the resolution of the text in these algorithms, while some algorithms inherently limit the size of the text to certain ranges.

2.4 Text-like texture segmentation in videos

Most video is often used to represent the moving scenes with a stream of image frames. Some movement information about the objects in the scene hide between continuous frames. Both the daily experience and research clues from psychology show that motion could be used to distinguish different physical object in vision system. Therefore, many researchers have tried to use movement information to improve the text segmentation in the video stream. This work is usually performed in two steps. The first step is to compute the motion in video. Pixels or blocks of pixels are specified in one frame, and then similar pixels and block of pixels are found in the next frames according to some certain measure. The vector from the position of these pixel or block in the first frame points to the next frame is used to represent the movement information. Second, this movement information can be used to improve the segmentation of the text in many different ways based on different assumptions that we make on text and its background.

Liernhart [34][33] has reported such a kind of media OCR system. First, the text segmentation step is employed to produce a binary image that depicts the text appearing in the video. Then, the



(a) Frame n

(b) Frame n+1

(c) Applying split & merge

(d) Applying size restriction

(e) Applying motion analysis

(f) Applying contrast analysis

FIG. 4 – *Video text process flow presented by Rainer Lienhart.*

standard OCR software packages are used directly to recognize the segmented text. The segmentation step consists of several processing steps. The color images are first processed with a region-based algorithm (split-merge) to segment the whole image into larger homogeneous regions. Then, the region images are binarized with local color contrast threshold. Some heuristics about the height and width of possible text regions are then applied to remove unlikely candidates. Moreover, the potential words or lines of the text are extracted through estimation of writing direction under the assumption that the text is aligned in horizontal line. It is also assumed that text appears in several consecutive frames, either on the same position or it is subject to linear movement. Text candidates that therefore do not appear in a suitable number of consecutive frames are discarded. The correspondence between several frames is computed by a simple region-matching algorithm and checked over five consecutive frames. The output of the text segmentation can be a binary image showing the extracted characters at their original location.

The advantage of this method is the ease of incorporating of the existing clean OCR software into the new system. For some applications, e.g. titles and credits as described in [33], the recognition rate ranging from 41% to 76%, which is lower than other methods discussed in the next subsections. Because there are little revioues work in this research direction, it is still difficult to say whether this kind of architecture can produce a satisfactory text recognition result.

Sato [31] presents an architecture to use the movement information in another way. The recognition algorithm performs video OCR on captions in the news videos, which consists of low resolution characters and widely varying complex background. In the first step, the algorithm roughly detects the text region as a horizontal rectangular structure of clustered sharp edges. The second step of the algorithm aims at improving the image quality. Two methods are used in this step. One method apply sub-pixel interpolation to obtain higher resolution images. The other method, called multi-frame integration, uses the movement information to enhance the foreground from the complex background. As a integrated OCR algorithm, the characters are extracted by four specialized line element filters and projection profile analysis. The algorithm keeps multiple segmentation results due to the difficulty of the character segmentation. In the last step, multiple segmentation results are used as inputs of the OCR system. The recognition is enhanced using dictionary based post processing. The algorithm assumes that the news captions are still and not always high contrast to the moving background. The pixel-based motion analysis is utilized mainly to improve the contrast.

There are also methods to treated the video as independent frames. For example, the algorithm presented in [2] employs this kind of architecture. The algorithm uses neural network to classify each input image frame into text pixel class and non-text pixel; class. After smoothing the classified image frame, the text pixels are extracted by applying binarization.

When the video are processed as independent frames, it almost can be regarded as still images. The two proposed motion analysis methods are based on pixel or rectangular block of pixels, while the pixel-based method assumes the text pixels are still and block-based method constrains that the movement of the text should be linear from top-to-down or left-to right.

2.5 Text Recognition in images and video

OCR system have been researched for many years. Some OCR systems can archive high recognition rate especially for the machine printed characters on the clean background. Since most of these systems ask for the input image as binary still image or still image that is easily binarlized, they can not be used to recognize the text in images or videos directly. Based on the text segmentation technologies. Presented in the last section, there are two ways to build the video OCR system. One method focus on improving the preprocessing in the clean OCR system with certain text detection, location and segmentation technologies to get more accurate text segmentation. When the proper parts of the text texture are extracted from the images completely, they are used as the input of the clean OCR system to produce the final text information. These proper parts of the text texture can be block of the texts, lines, words or single characters, which depend on the different application systems. The architecture used in [36][38][37] is a good example. The input image is first cleaned up with a four-step processes,

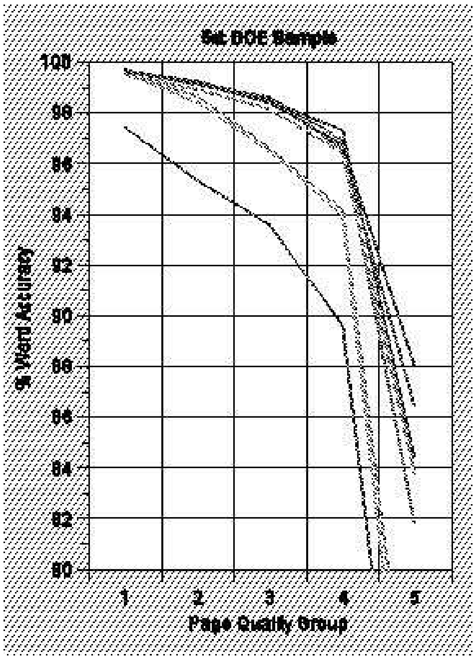


FIG. 5 – *typical commercial OCR software performance on page quality*

and then commercial OCR software is employed to do the recognition. In this way, the new system can utilize existing OCR technologies. The disadvantage of this method is that recognition is limited by the performance of existing OCR technologies. Although text segmentation processes can clean up the images in some degree, the segmented images output by the process are still unsuitable for traditional OCR. They are usually noisy, have low resolution, and the character display transformation. The final recognition performance of the system has to depend on the noise adaptation of the employed clean OCR technologies. Figure x shows the performance of some typical commercial OCR software on resolution and page quality.

The statistic data is come from the ISRI 1995 annual test of OCR accuracy [18]. The different curves in the figure represent different software. The page quality groups in (b) , ranging from 1 to 5, represent the quality of the pages from high to low. Here the page quality only measures the image with clean background.

The other way to build dirty OCR system is to develop new character recognition technologies to adapt the heavy noise cause by the complex background. Sato's algorithm is a good example. In the paper [32], a comparison is made between the new developed recognition algorithm and the conventional recognition software. Being applied on the news video stream, the algorithm developed by the authors reach the recognition rate from 76.2% to 89.8%, which is much higher than the recognition rate, ranging from 38.9% to 53.2%, of the conventional OCR.

3 Detailed Technologies

A lot of image & video processing methods are used in text detection and recognition in images and videos. In this section we will discuss these methods in two stages: text segmentation and text recognition.

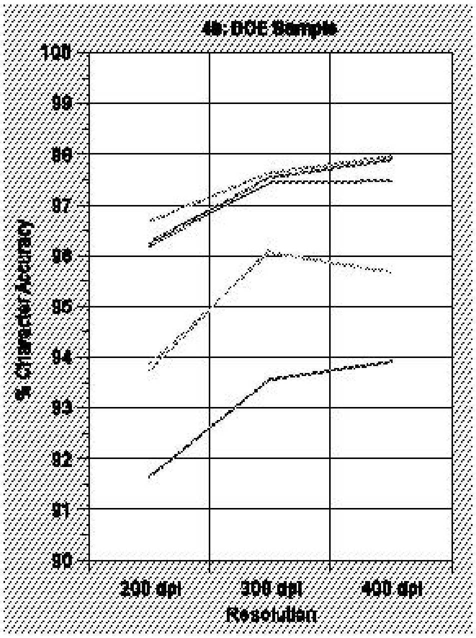


FIG. 6 – *typical commercial OCR software performance on image resolution*

3.1 Text Segmentation

As we discussed in the last section, the aim of the text segmentation is to classify the pixels of the images or video frames into two classes. Ideally, one class consists of the pixels of the text. The other class contains the pixels of the non-text.

3.1.1 Segmentation based on image statistic property

Binarization

A large number of binarization methods have been proposed in the literature. They have usually been developed for specific applications (e.g. postal address, checks), that allow strong constraints on document structure (e.g. layout, character size).

A typical binariation method uses a single threshold value which is applied over the whole image [6][7], also called global-threshold method. These can be classified into non-parametric, parametric, and other methods.

Non-parametric methods: A non-parametric method has been described in [8]. This method calculates the ratio of between-class and within-class covariance for each potential threshold, where the two classes represent the foreground and background pixels. The purpose is to find the threshold that maximizes this ratio which is then chosen as the threshold. A similar method that aims to maximize the entropy of the two classes has been described in [9]. Another method is based on moment preservation [10]. It finds the threshold which best preserves the moment statistics in the resulting binary images as compared to the original grey-level image. The initial moments are calculated from the intensity histogram.

Parametric methods: A parametric thresholding method based on minimum error thresholding has been reported in [11]. It is based on pattern recognition techniques where the foreground and background intensity distributions are modeled as normal probability density functions. The threshold is chosen so that the classification error between the two classes is minimized. It has been found that error rate with this method are lower than those with non-parametric methods [13].

Other methods: A binarisation method for printed address blocks can be found in [14]. The tech-

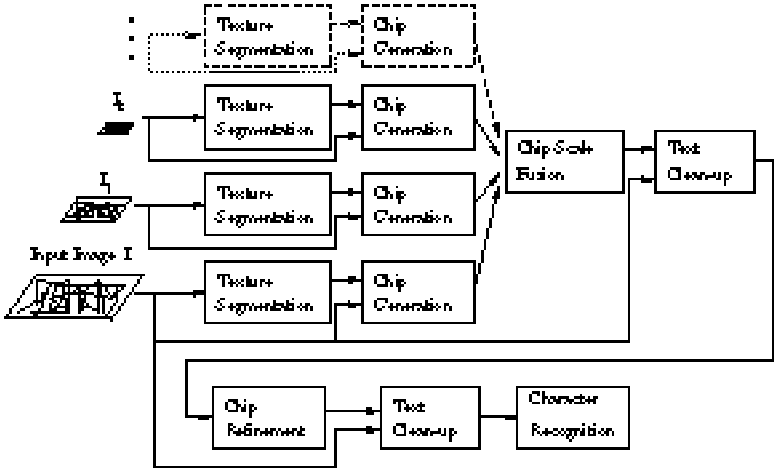


FIG. 7 – *Different scale text processing method presented by Victor Wu.*

niques consists of three steps: first, candidate thresholds are produced through iterative use of Otsus method [8]. Texture features associated with each candidate threshold are then extracted from the run-length histogram of the accordingly binarised image. Finally, the desired threshold is selected so that desirable document text features are preserved. The methods has been shown to achiever good results on a large number of unconstrained mail addresses.

Adaptive thresholding: One possibility to the threshold of non-homogeneous regions is the application of adaptive thresholding techniques. These methods usually analyses local windows across the image on which adaptive thresholding is performed. The main problem of these methods is the choice of the window size. The window should be large enough to include a representative number of pixels in the window but not so large as to average over non-uniform background intensities. Different techniques were evaluated for documents such as check's with background pictures, shadows, highlights, smear, and smudge [15]. An evaluation of a large number of different methods has been reported in [16]. It shows that the performance of these methods are application dependent.

Thresholding methods usually achieve good results on images with homogeneous background. However, when applied to non-homogeneous background, as is often the case for text in images and video, results are often unsatisfactory, which has been noted by several researchers [7][16].

Region analysis

Scale-space method: [36] presented a scale-space algorithm to cluster the image pixels into regions. It employs three second order derivatives of Gaussian's at three different scales as filters to compute 9-dimension of feature vectors for each pixel in the image. The text string regions are decided by clustering the 9-dimension feature vectors into interested class. In other word, one of the nine classes is regarded as text string region class. The scale approach used in the algorithm is also properly to solve the variation of the text size.

[28][27] proposed a method to located text regions based on edge detection, which is called the spatial variance method. It computes the local spatial variance along each horizontal line over the whole input gray-scale image. (in this method, the text is assumed to be aligned horizontally.) Then, an edge detector is used to find the significant horizontal edges. Finally, the edges with opposite directions are paired into lower and upper boundaries of a text line. For the spatial variance method, the main disadvantage is that it sometimes cuts the extending characters, such as g, p, into halves.

[27] also proposed a method to locate the text in color images, which is called connected component method. This method first reduces the RGB color values from 24 bits to 6 bits by retaining the high 2 bits for R, G, and B value. The color reduced images are decomposed into mono-color components.

The color, which belongs to the largest number of pixels in the image, is regarded as the background. The color components are then connected using heuristics into the text line. At the next stage, the algorithm uses OCR system to identify the text in the connected component. The identified connected component will be extended to find lost characters in the neighborhood. This algorithm may use some shape information of the characters, such as block adjacency graph [27] to to do connected component analysis. It works for both gray scale images and color images, but asks for the characters to have a distinct color from the background. The main shortcoming of the connected component algorithm is that it can not recognize characters which are connected. In [28], the authors present a method to merge the connected component method and the spatial variance method. The text candidates are first located with spatial variance, and then segmented with connected component. Furthermore, the color of the text is determined inside the candidate box, and the text components are located. Finilly, the identified connected component will be extended to find the lost characters in the neighborhood.

3.1.2 Segmentation based on motion analysis

The motions in the scene are cause by movements of the objects in the scene and changing parameters of the camera system. Motion analysis can offer important clues for segmentation. For example, if we assume that all the objects in the scene are rigid objects, then the parts of the frames with different movement should be segmented into different objects. Of course, the movement, here, refers to 3-D movement. Under certain assumptions, such as the rigid object assumption, linear movement assumption, even the 2-D movement information is helpful for segmentation.

Motion analysis commonly consists of three steps. First, some feature elements are located through the continuous frames. These feature element can be pixels, features, or block of the image. Second, the correspondence of the feature elements are found between the frames. Finilly, the motion vectors are computed for the feature elements and used in some further analysis. Two such the methods have been used in the text segmentation in videos.

One is presented in [34]. This method assumes that the movement of the text is limited as 2-D rigid translation movement. The feature elements in this algorithm are rectangular blocks of the image, which are assumed to contain text. The process of correspondence employs the block-based motion estimation algorithm with MAD (mean absolute difference) criterion within constant size of searching area. After the motion vectors are computed, the blocks are discarded if they do not have the correspondences or the average gray tone intensities of these blocks show significant difference with their correspond blocks. Block-based motion estimation algorithms have been researched for many years and applied on varied application areas, for example video compression. One of the advantages of this block-based motion analysis is that there are a lot of fast block-based motion estimation algorithms, e. g. three-step searching and pyramid searching. One of the obstacles to applying block-based motion analysis is how to specify the block, which contains exactly a word or text string which always keeps moving together. In other words, the block-based motion analysis can only be used after the rough word or text string region has been detected. The other disadvantage of block-based motion analysis is that the pixels of the text usually can not cover solidly within the block. There are always many areas between the characters or in side of the character that move differently with the text. If they are regarded as a integrated part at what block-based motion estimation does, there will be many errors in the motion estimation results.

The other method of motion analysis is described in [32]. This algorithm assumes that the text is always stationary while the background is moving. It also assumes that the gray value of the text is white. According to these two assumptions, the feature element, here the single pixel, is valued with the minimum gray value of the pixel at the same position through multiple frames. This algorithm is easy to implement and run very fast. Unfortunately, it can only be used when the text is stationary and the background is moving. The constrain about the gray value of the text is another obstacle to extending the algorithm to wider applications. Also the assumption that background is darker than text is not always true.

Different forms of movements exist in different types of videos, while different motion analysis can be used to enhance the text detection and segmentation in videos. The two simple cases are either the text is stationary and the background is moving or vise versa when there is only 2-D translation movement. This kind of movement can be detected with the variation of the gray values. It should be pay attention to the motion analysis results when the contents of the text change in between the frames. Some other constraints, for example the linear movement, could be exploited for disregarding the non-text areas. More complex movements of both text and background are still difficult to be used in text segmentation.

3.1.3 Segmentation with shape and texture modeling

If the image is very complex, the statistic information of text pixels and background is not enough for the text segmentation. Therefore, people use geometrical information of the shape of the characters to enhance the text segmentation. Methods proposed in the literature can be classified into line element detection approaches, bottom-up approaches, and texture based techniques.

Line element detection methods: A typical line element detection method is based on run length smoothing algorithm [19][20][21]. The algorithm requires a binary image and replaces every string of contiguous 0s of less than a predefined value by a string of 1s of the same length. The algorithm is usually applied in horizontal and vertical direction and the results are used to classify regions into horizontal lines, vertical lines, text, and images. The drawback of this method is that several thresholds have to be chosen and that text embedded in images is not detected [22]. Another line element detection method is based on specialized filters [32]. The algorithm employs four filters in the final binarization step, which correspond to the line elements of the character: vertical, horizontal, left diagonal, and right diagonal. The size of the filter is defined to include only a line element of the characters. Integration of the the four filters, the algorithm can reduce the effect of the complex backgrounds.

Bottom-Up methods: Bottom up method typically work by grouping pixels as connected components which are merged into successively larger components until the whole document has been processed [23]. The algorithm performs well if certain requirements are fulfilled, including character size, inter-line spacing, inter-character spacing, and resolution.

Texture segmentation based methods: The third category is based on texture segmentation. This assumes that text and non-text regions belong to two different texture classes that can be separated by texture analysis methods. A small number of two-dimensional Gabor filters have been used in [26][25] for document segmentation. The method has been shown to work well for newspaper images, which share low resolution as video but have large block of text and stable background.

3.1.4 Segmentation with neutral network

A text segmentation method that uses a neural net to classify the output of wavelets has been described in [30]. A neural net is trained with the outputs of multi-scale wavelets extracted from image regions belonging to foreground and background regions. Segmentation experiments on a few images have shown good results. However, the neural net was trained on the same images as it was tested on, it is therefore difficult to judge the generalization ability of the method.

The other two neural network based approaches for text segmentation have been described individually in [28] [2]. in [28], a MLP is trained to discriminate three main categories: half-tone, background, and text and line-drawing regions. The input layer of the network is trained with the grey-scale pixel values inside a 7×7 window. In [2], a three layer neural network is used to classify grey-level pixles into text pixels, and non-text pixles. As we introduced Jain's algorithm in [25], many algorithms employ neutral networks as tools to segment the text from media.

3.2 Text recognition technologies

There are two ways to fulfill the text recognition (i) use of a commercially available OCR system, (ii) develop special OCR technologies for text recognition in images and videos. For superimposed

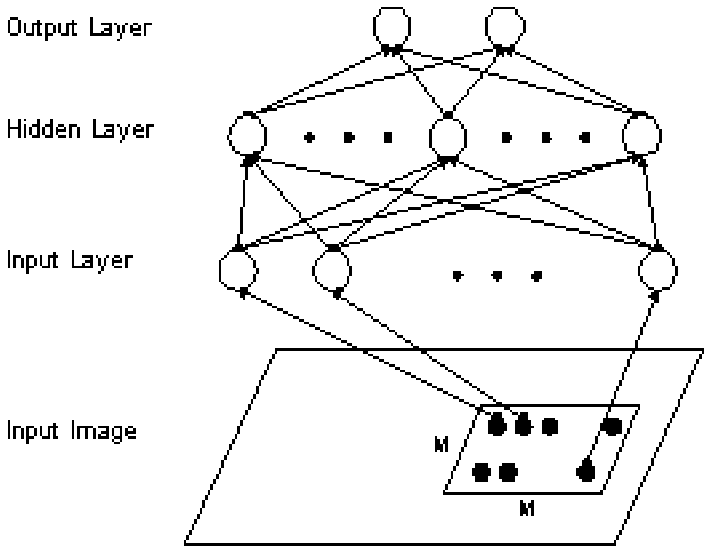


FIG. 8 – *An example to segment the text with NN*

text, which is more controlled, the application of a public OCR system to be appropriate. Many text enhancement processes have been employed by researchers to apply commercial OCR system to media text recognition. Of course, some processes to enhance the text video frame are necessary.

3.2.1 Text enhancement through interpolation

The size of some texts, especially in video, is too small (less than 10 x10) to be used as the input of the commercial OCR system. In [31], a linear interpolation has been utilized to improve the resolution of these texts. Each text area is magnified four times in both x and y directions. The sub-pixels are interpolated by linear function using neighbor pixel values of the original image weighted by the specified distances. By using this linear interpolation method, the low resolution text regions can be extended to enough higher resolution images, which can be adopted by commercial OCR system. However, the linear interpolation is noise sensitive. An alternative way is use the bilinear interpolation instead of linear interpolation [1].

3.2.2 Text Enhancement through multi-frame process

Many evidences show that the same text string often spans several frames in a video stream. This multi-frame information has been explored to filter out the noise or enhance the contrast between text and background. In the above section, we have mentioned two methods about utilizing movement information to improve the text segmentation, which are presented individually by Leinhart and Sato. In Leinhart's paper, the motion vectors of the text candidate blocks are used to filter out the non-text region by checking whether the blocks have linear H/V movements. This method employs the MAD as the criterion to match the correspondent blocks in different frames.

$$MAD(d_1,d_2) = \frac{1}{|R|} \cdot \sum_{(x,y)\in R} |g(x,y) - g(x+d_1,y+d_2)|, \quad |d_1|, |d_2| \leq search\,range$$

R specifies the block of text region. $g(x,y)$ denotes the gray image.

The method adapts the movements of both the text and the background. The disadvantage of this method is that it can only be used to select the text block candidates, but not to enhance the text

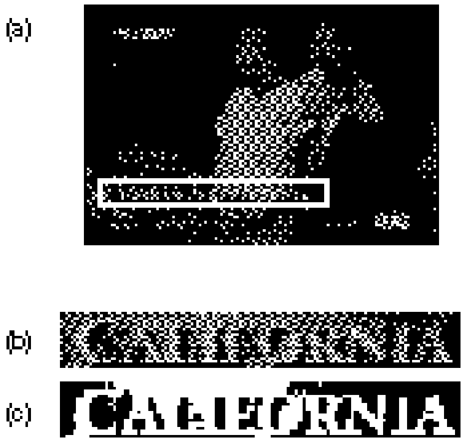


FIG. 9 – *Sato's method: (a) the original image. (b) Magnified image of character. (c) Binary image.*

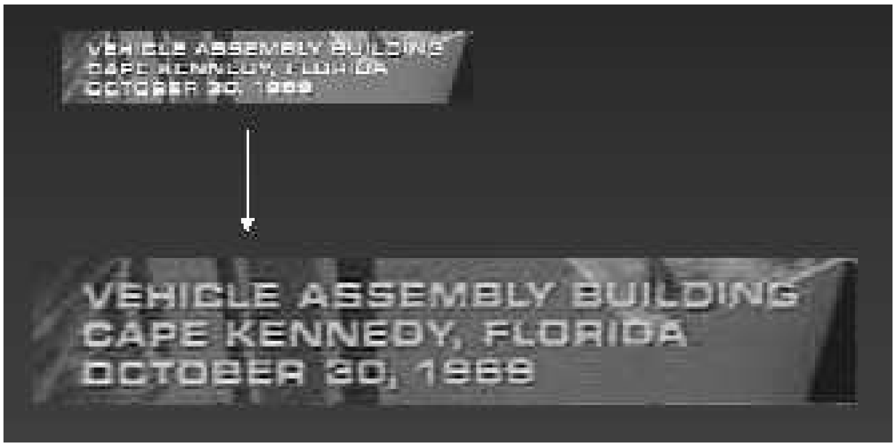


FIG. 10 – *Bilinear interpolation text enhancement presented by Huiping Li.*

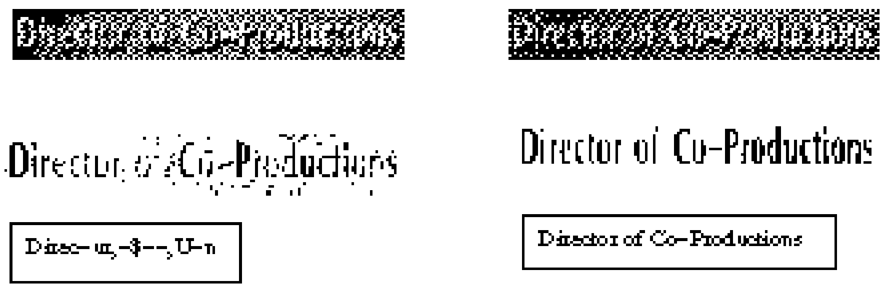


FIG. 11 – *Multi-frame text enhancement example presented by Huiping Li. The images on the left side show the OCR result in single frame. The images on the right side show the OCR result after averaging over multiple frames.*

pixels in each block. Sato present the algorithm which can enhance the text pixels by multi-frame integration. The value of the pixel at the same position within n continue frames is decided by the following equation:

$$L_m(x,y) = min\,(L_i(x,y), L_{i+1}(x,y), \ldots, L_{i+n}(x,y))$$

$L_m(x,y)$ is the enhance value of the pixel, where (x,y) indicates the position of the pixel. However, this algorithm only process the frames with static text.

In [1], a feature point based method is presented to enhance the text with the multi-frame information. Some feature points are selected in each candidate text block and matched through the sequence of the frames under the pure translation constrain. These feature points matching results carry the information about the motion vector of the text block. Furthermore, the text pixels are enhanced by smoothing out the noises within the corresponding text block sequence. This method assumes that the noise $N(t)$ is a Gaussian random process satisfying $N(t) \sim N(0,\sigma^2)$, and the pixels of both text and background in the same text block keep the same value through the corresponding sequence. the method offers a way to enhance the text pixels when the text is moving. The challenge of this method is how to select and track the feature points to obtain accurate corresponding.

3.2.3 Overcome the degradation

Edge degradation of the binarization result of the text image is an obstacle for the edge-based text recognition. Edge degradation is caused by low spatial resolution of the image or video frame and misjudging the background as a part of text or vise versa. The considered effective edge-based features, for example, LDC (Local Direction Contributivity) [41] and the Chaincode feature [24]. As an improvement, a feature is proposed in [17], which is called (Double-region LDC) WLDC. Both of the LDC in foreground and background are extracted and combined together to produce the proposed feature. The experiment results show that with certain classifier the WLDC can improve the Telop character recognition rate.

4 Conclusions and Discussion

The research on media text detection and recognition aims extracting and recognizing the text in arbitrary images and videos, which suffered by the text's variability of size, font, color, orientation, style, alignment, (even within words), and partially occluded, opaque. Proposed research work introduced in this survey have shown many merits in solving the problems exist in different applications.

4.1 Discussion about technologies

Color analysis In many images, the text color is chosen for high readability in color images, without considering it's readability in gray-level converted form. This can result in low contrast between characters and background when converted to gray-scale images Some systems are based on gray-scale images [31], others [35] have observed improved results by the use of color information. Different color spaces like R-G-B (Read, Green, Blue) and I-H-S (Intensity, Hue, Saturation) may lead to different results for the detection and segmentation of text.

Proposed methods for color based segmentation are often based on the segmentation of connected regions of uniform color. These techniques usually assume that the text color within a text region is constant, which is often the case. However, text in video images is often of lower quality that can result in varying text color within a text block. Within one image, different text blocks can appear in different colors. It is therefore important to consider several color regions as potential text candidates [27], in order to detect all text regions in the image. This issue of color variability and its effect on color-based segmentation methods need to be investigated.

Texture analysis The use of texture analysis has been shown to lead to good results in text detection. These methods are typically based on multi-channel filtering or on spatial analysis. One difficulty with texture based detection is that the detection results are not always very accurate, i.e. the bounding boxes found by the algorithm do not always completely overlap with the actual text. Some proper sort of post-processing to determine the the text boundaries more accurately.

Texture filtering methods do usually only consider small local regions of the image. The outputs therefore need to be further processed by additional filters and clustering methods. A filtering method that considers textual properties of one or several entire characters might be a more promising approach. The difficulty with this method, however, is to design a technique which is general enough to detect text regardless of the character property (size, font, etc.) but specific enough to discriminate between text and between non-text.

Geometric analysis Most current systems use some kind of geometric heuristics about the minimum and maxima of text regions as well as about the ratio of height to width. Although these have lead to good results, it is not clear how these are determined and not known how they generalize for a larger and broader set of test images. A more data-driven or probabilistic approach might be more desirable to incorporate geometric information. This could for example be obtained from a representative set of images.

Motion analysis Motion analysis of video images offers several possibilities to enhance text detection. The fact that text appears in several consecutive images can be used to disregard possible text candidates that do not appear in a minimum number of images. Constraints of linear movement and uniform background can be exploited for the detection of moving text. Moving background can be used for the segmentation of stationary text. In addition, temporal information can be used to evaluate a number of different segmentation and recognition alternatives, by applying these techniques on several consecutive frames and choosing the most likely alternative.

Evidence Combination Several information sources can be used for text detection (e.g. color, texture, geometry, motion) where each method will lead to a different performance depending on the test material. It is usually advantageous to combine different information sources in order to improve the performance of individual methods and to make them invariant to different test conditions. The problem to combine the outputs of alternative approaches based on probabilistic or other technologies need to be carefully investigated.

All the presented work address the text recgontion in image and video in a way with three steps: text location, segmentation, and recognition. In many cases, good text locations and segmentations require the detail information of the shapes of the text characters. Unfortunatly, until now the shape information of the characters can be obtained only when the characters are well segmented from the background. This chicken-and-egg problem seems to be the main obstacle to improve recognition rate.

5 References

Références

[1] Huiping Li & David doermann, "Text enhancement in digital video using multiple frame integration", ACM Multimedia 1999.

[2] Ki-Young Jeong, Keechul Jung, & Hang Joon Kim, "Neural network-based text location for news video indexing", ACM Multimedia 1999.

[3] M. Bokser. Omnidocument technologies. *Proc. IEEE*, 80(7):1066–1078, July 1992.

[4] S. V. Rice, F. R. Jenkins, and T. A. Nartker. OCR accuracy: UNLV's fifth annual test. *INFORM*, 10(8), September 1996.

[5] L. O'Gorman and R. Kasturi. *Document Image Analysis*. IEEE Computer Society Press, Los Alamitos, 1995.

[6] C. A. Glasbey. An analysis of histogram-based thresholding algorithms. *CVGIP: Graphical Models and Image Processing*, 55(6):532–537, 1993.

[7] L. O'Gorman. Binarization and multi-thresholding of document images using connectivity. *Computer Vision, Graphics and Image Processing*, 56(6):494–506, 1994.

[8] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979.

[9] T. Pun. Entropic thresholding: A new approach. *Computer Vision, Graphics and Image Processing*, 16(3):210–239, 1981.

[10] W. H. Tsai. Moment-preserving thresholding: A new approach. *Computer Vision, Graphics, and Image Processing*, 29:377–393, 1985.

[11] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[12] J. Ohya, A. Shio, and S. Aksmatsu. Recognition characters in scene images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(2):214–220, 1994.

[13] A. S. Abutaleb. Automatic thresholding of gray-level pictures using two-dimensional entropy. *Computer Vision, Graphics and Image Processing*, 47(1):22–32, 1989.

[14] Y. Liu and Sargur N. Srihari. Document image binarization based on texture features. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(5):540–544, May 1997.

[15] M. Kamel and A. Zhao. Extraction of binary character/graphics images from grayscale document images. *Computer Vision, Graphics and Image Processing*, 55(3):203–217, 1993.

[16] Ø . D. Trier and A. K. Jain. Goal-directed evaluation of binarization methods. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(12):1191–1201, 1995.

[17] M. Mori, S. Kurakake, T. Sugimura. Robust Telop character recognition in video for content-based retrieval.. *Proc. of ICDAR'99*, pp:13–16, 1999.

[18] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fourth annual test of OCR accuracy. ISRI annual research report, 1995.

[19] F. M. Wahl, K. Y. Wong, and R. G. Casey. Block segmentation and text extraction in mixed text/image documents. *Computer Graphics and Image Processing*, 20:375–390, 1982.

[20] K. Y. Wong, R. G. Casey, and F. M. Wahl. Document analysis system. *IBM Journal of Research and Development*, 26(6):647–656, 1982.

[21] D. Wang and S. N. Srihari. Classification of newspaper image blocks using texture analysis. *Computer Vision, Graphics and Image Processing*, 47:327–352, 1989.

[22] G. Nadler. A survey of document segmentation and coding techniques. *Computer Vision, Graphics and Image Processing*, 28:240–262, 1984.

[23] L. A. Fletcher and R. Kasturi. A robust algorithm for text string seperation from mixed text/graphics images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9(1):149–153, 1987.

[24] F. Kimura, K. Takashima, S. Tsuruoka, and Y. Miyake. Modified quadratic discriminant functions and the application to chinese character recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(6):910–918, 1988.

[25] A. K. Jain and S. Bhattacharjee. Address block location on envelopes using gabor filters. *Pattern Recognition*, 25(12):1459–1477, 1992.

[26] A. K. Jain and S. Bhattacharjee. Text segmentation using gabor filters for automatic document processing. *Machine Vision and Applications*, 5:169–184, 1992.

[27] A. K. Jain and B. Yu. Automatic text localisation in images and video frames. *Pattern Recognition*, 31(12):2055–2076, 1998.

[28] A. K. Jain and Y. Zhong. Page segmentation using texture analysis. *Pattern Recognition*, 29(5):743–770, 1996.

[29] R. Jain. Visual information management. *Communications of the ACM*, 40(12):31–32, December 1997.

[30] K. Etemad, D. Doermann, and R. Chellapa. Multiscale segmentation of unstructured document pages using soft decision integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):92–96, 1997.

[31] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith. Video ocr for digital news archives. In *IEEE Workshop on Content Based Access of Image and Video Databases*, Bombay, January 1998.

[32] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh. Video OCR: indexing digital news libraries by recognition of superimposed caption. In *ACM Multimedia System Special Issue on Video Libraries*, Feb. 1998.

[33] R. Lienhart and W. Effelsberg. Automatic text segmentation and text recognition for video indexing. Technical Report TR-98-009, University of Mannheim, Mannheim, 1998.

[34] R. Lienhart. Automatic text recognition in digital videos. In *Proc. SPIE, Image and Video Processing IV*, January 1996.

[35] R. Lienhart. Indexing and retrieval of digital video sequences based on automatic text recognition. In *Proc. 4th ACM International Multimedia Conference*, Boston, November 1996.

[36] V. Wu, R. Manmatha, and E. M. Riseman. Finding text in images. In *Proc. ACM Int. Conf. Digital Libraries*, 1997.

[37] V. Wu, R. Manmatha, and E. M. Riseman. Textfinder: An automatic system to detect and recognize text in images. Technical report, University of Massachusetts, Amherst, MA, 1997.

[38] V. Wu, R. Manmatha, and E. M. Riseman. Textfinder: An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1224–1229, 1999.

[39] V. Wu and R. Manmatha. Document image clean-up and binarization. In *Proc. SPIE Symposium on Electronic Imaging*, 1998.

[40] Y. Zhong, K. Karu, and A. K. Jain. Locating text in complex color images. *Pattern Recognition*, 28(10):1523–1536, 1995.

[41] T. Akiyama and N. Hagita. Automated entry system for printed documents. *Pattern Recognition*, 23(11):1141–1154, 1990.